5
Washback

Dianne Wall

Introduction

It is now well accepted that tests can have important consequences – for students, whose future may be determined by their test results, and for their teachers, whose self-esteem, reputation and even career progression may be affected by how successful they are at preparing their students to cope with test requirements. If the test is 'high stakes', defined by Madaus as a test 'whose results are seen – rightly or wrongly – by students, teachers, administrators, parents, or the general public, as being used to make important decisions that immediately and directly affect them' (Valette, 1988: 87), a phenomenon known as 'washback' may occur. Washback is the term that is used when students and teachers 'do things *they would not necessarily otherwise do* because of the test' (Alderson and Wall, 1993: 117). They might, for example, pay more attention to certain parts of the teaching syllabus at the expense of other parts because they believe these will be emphasised on the test. They might practise certain types of questions rather than others for the same reasons. The washback of a test

> can either be positive or negative to the extent that it either promotes or impedes the accomplishment of educational goals held by learners and/or programme personnel.
>
> *(Bailey, 1996: 268)*

There has been growing interest in washback over the last two decades, for theoretical, practical and political reasons. Messick (1996) claimed that washback is a particular instance of the consequential aspect of construct validity, which suggests a corollary that investigating washback and other consequences is a crucial step in the process of test validation. The practical and political interest stems from the use that policy-makers make of high-stakes tests to influence educational practices. Educational advisors such as Heyneman and Ransom argue that tests 'can be a powerful, low-cost means of influencing the quality of what teachers teach and what students learn at school' (Valette, 1992: 105). This view is shared by educators such as Popham (1987), who proclaimed the advantages of 'measurement-driven instruction', and language educators such as Pearson (1988), who used the term 'lever for change' to describe the influence he hoped a new national test of English would have on language teaching. Opponents to this view include Madaus, who claimed that using high-stakes testing to effect change leads to 'teaching to the test'

and the use of past test papers as teaching material, and that it 'transfers control over the curriculum to the agency which sets or controls the exam' (Valette, 1988: 97).

The idea of washback takes on more complexity when we consider not only whether the effect of tests are positive or negative but also whether they are immediate or delayed, direct or indirect, or apparent or not visible – e.g. changes in attitude that do not manifest themselves in overt behaviour (Henrichsen, 1989: 80). There are also methodological challenges in determining whether washback has occurred, such as deciding on procedures to establish whether the classroom practice that appears after the introduction or revision of a test is actually test washback, or rather the effect of other factors in the educational context.

Historical perspectives

There are many accounts of the use of high-stakes testing in education and other realms of public life. Eckstein and Noah (1993: 5–17) discuss a number of functions that such tests have served in society: ending the monopoly over government jobs held by the privileged classes (e.g. competitive examinations in China during the Han Dynasty), checking patronage and corruption (the Indian Civil Service examination in nineteenth-century Britain), encouraging 'higher levels of competence and knowledge' (entry examinations to the professions in France and Germany), allocating sparse places in higher education (university entrance examinations in Japan), and measuring and improving the effectiveness of teachers and schools (the 'Payment by Results' system established in Britain in the 1860s, where state funding of schools was partly determined by the results students received in tests administered by school inspectors). It is not difficult to imagine the influence that these tests might have had on the learning goals and methods of the candidates preparing for them. There is little empirical evidence available, however, to provide a link between these tests and the teaching and learning that are said to have resulted from them.

There was considerable discussion of the effects of high-stakes testing in educational settings in the twentieth century. Fredericksen and Collins introduced the notion of 'systemic validity', which occurred when a test

> induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure.
>
> *(Valette, 1989: 27)*

These changes were not to be seen as automatic, however. Improvements in educational standards depended not only on well-constructed tests representing the full range of educational objectives, but also on a clear articulation and exemplification of the standard desired, sensitive training of markers, and ample opportunity for teachers and students to understand and practise using the criteria for evaluating performance. Airasian (1988) was cautious in his views of whether tests could bring about changes in teaching and learning, emphasising the need to consider the level of cognitive skills assessed in the test and the likelihood that teachers could successfully train students in high-level operations such as reasoning, problem-solving and critical thinking. Educators such as Haladyna *et al.* (1991) signalled some of the dangers of using tests to change standards, including threats to validity caused by test preparation practices which 'increase or decrease test performance without connection to the construct represented by the test' (4). They ranked different types of preparation activity according on a scale from ethical to highly unethical. An example of an ethical practice was motivating students to study by discussing how important the test was to them. An example of unethical practice was using material in the classroom that

was very similar to the material used in the test (see Mehrens and Kaminsky, 1989, for further discussion of the ethics of test preparation).

It was not until the early 1990s that discussions regarding the washback of high-stakes testing began to appear in the field of language testing. Until this time there were general claims about the influence of tests on the curriculum, but little discussion of whether this influence was real or imagined. Alderson and Wall (1993) provided the first critical analysis of the notion of washback. They explained that washback was not the same as the general pressure caused by important tests – for example, the feeling that one should spend more time studying. The washback of a test was specific to that test alone, manifesting itself in decisions about how much attention to pay to certain aspects of the domain in question (e.g. the teaching syllabus), depending on the importance given to these aspects in the test. They proposed a number of 'washback hypotheses' to illustrate the types of influences that a test could conceivably have: on what teachers taught and learners learned; on how teaching and learning took place; on the rate, sequence, degree and depth of teaching and learning; and on the attitudes of teachers and learners. They stressed how important it was for test designers to make explicit the type of washback they intended to create with their tests, and for researchers to be explicit about the types of activity they were investigating in their attempts to discover whether washback had occurred. Alderson and Wall further argued that there was a need to expand the range of research techniques used to study washback, and to draw on the insights gained in other fields of research (such as innovation and motivation) in order to not only describe washback but also to account for its nature.

Other early contributions to the understanding of washback came from Hughes (1994) and Bailey (1996, 1999). Hughes introduced an important distinction between washback on the participants in an educational system (e.g. teachers, learners, administrators, materials writers, curriculum designers), on processes (the types of thinking and activities the participants engaged in) and on products (for example, the amount and quality of learning that resulted). Bailey, among her other contributions, created a 'basic model' to illustrate the mechanism by which washback developed. She also proposed a series of questions to ask of any external examination (one which was not created within the institution where it was being administered). These questions probed the participants' understanding of the purpose of the test and how the results were to be used, the theoretical bases of the test, the use of authentic texts and tasks, the manner in which test results were provided, and other features she believed could influence the appearance and nature of washback.

The next important milestone appeared in 1996, with the publication of a special issue of *Language Testing* devoted to washback. Issue 13: 3 included two discussions of the nature of washback and four case studies. The first discussion was by Messick (1996), who first reviewed the notion of a unified version of construct validity, and then argued that washback was one (but not the only) manifestation of the consequential aspect of validity. It was important to investigate washback when attempting to establish the validity of a test. Messick also warned that just because certain features appeared in the context where a test was operating, this did not mean that those features constituted washback. In his words

> it is problematic to claim evidence of test washback if a logical or evidential link cannot be forged between the teaching or learning outcomes and the test properties thought to influence them.
>
> *(Messick, 1996: 247)*

If a link could be found between the test and the teaching or learning outcomes, and the outcomes were not considered desirable then it would be necessary to analyse the test design and

see whether it required adjustments. It was preferable, however, to concentrate on creating desirable washback to begin with,

> … rather than seeking washback as a sign of test validity, seek validity by design as a likely basis for washback.

Messick advocated incorporating authenticity and directness in the test design, and to minimise 'construct under-representation' and 'construct-irrelevant difficulty' (252).

The second discussion of the nature of washback was by Bailey (1996), who provided a comprehensive review of the literature on washback up to that point, and presented the model of washback referred to above.

Four case studies provided the type of empirical evidence for washback that was often missing in publications before the 1990s. Alderson and Hamp-Lyons (1996) used teacher and student interviews and classroom observations to investigate the washback of the Test of English as a Foreign Language (TOEFL) on teaching in a private language school in the United States. They followed two teachers as they conducted 'normal' language lessons and TOEFL preparation lessons, and found that although there were differences between the two types of lessons for each teacher, there were differences that were at least as great between the two teachers themselves. Watanabe (1996) also designed a comparative study, with the aim of determining whether there was a relationship between university entrance examinations in Japan (which are very high stakes) and the amount of grammar-translation teaching taking place in private examination preparation centres. He carried out observations and interviews with two teachers who were teaching preparation classes for one examination which emphasised grammar-translation and for one which did not. He predicted that both teachers would include more grammar-translation teaching in the first type of class than in the second. He found, however, that while one teacher seemed to be influenced by the type of examination he was preparing his students for, the other explained grammar and employed translation no matter which type of examination he was dealing with. Both Alderson and Hamp-Lyons, and Watanabe concluded that tests do not affect all teachers in the same way: that the personal characteristics of the teachers, including their beliefs and attitudes, are key factors in how they react to test pressures.

Shohamy *et al.* (1996) and Wall (1996) investigated the washback of high-stakes tests on teaching in state secondary schools. Shohamy *et al.* compared the effects of two tests in Israel, a test of English as a Foreign Language (EFL) and a test of Arabic as a Second Language (ASL), both when they were introduced and after they had been in place for some time. They found that while the washback of the EFL test increased over the years, the washback of the ASL test decreased 'to the point where it has no effect: no special teaching activities are introduced in preparation for the test, no special time is allotted, no new teaching materials have been developed, awareness of the test is minimal … ' (Shohamy *et al.*, 1996). The researchers attributed these changes to a variety of factors, including the status of the two languages within the country, the purposes of the tests, the test formats that were used and the skills that were tested.

Wall (1996) reported on the washback of a major EFL test on secondary school teaching in Sri Lanka. Her study involved repeated observations at many schools over a two-year period, and in-depth interviews with the teachers whose classes were observed. She found that the test in question influenced the content of the classes in both positive and negative ways, but it had no influence on the methods the teachers used to deliver this content. Wall presented a number of reasons for these findings, relating not only to the test itself but also to other factors in the educational and social setting. These included a lack of understanding on many teachers' part of the test construct, a lack of official feedback to teachers about their students' test performance,

and inadequate teacher support systems. Wall's analysis was underpinned by insights from innovation theory (e.g. Fullan and Stiegelbauer, 1991).

Interest in washback grew quickly in the 1990s, with explorations into many dimensions of the phenomenon and the use of many frameworks and methods to gather and analyse data. The next decade witnessed a further growth in the number of investigations into washback and into the wider educational and social impact of high-stakes testing.

The second important collection of washback studies was edited by Cheng and Watanabe (2004). This volume contained three comprehensive reviews of different aspects of the topic: the impact of testing on teaching and learning (Cheng and Curtis, 2004), the methods (mainly qualitative) used in washback studies (Watanabe, 2004) and the relationship between washback and curricular innovation (Andrews, 2004). It also contained eight studies of the washback of different kinds of assessment (school tests, university entrance tests, tests for immigration purposes, international proficiency tests) in different education settings (the United States, New Zealand, Australia, Japan, Hong Kong, China and Israel) and on different aspects of teaching, including the design of course books to support test preparation.

A number of monographs on washback and other forms of impact have also been published. Cheng (2005) investigated how teachers and students in Hong Kong secondary schools reacted to changes in the English component of the Hong Kong Certificate of Education. Wall (2005) developed further her ideas concerning the relevance of innovation theory to washback studies (as noted in Wall, 1996, reviewed above) and provided a more complete treatment of examination reform in Sri Lanka. Green (2007) investigated the effects of the Academic Writing Module of the International Language Testing System (IELTS) examination on preparation for academic study. Hawkey (2006) presented two separate studies: one into the impact of the IELTS, and one into the impact of a foreign language education improvement programme in the state school system in Italy (the 'Progetto Lingue 2000'). These volumes, along with the case studies presented in the Cheng and Watanabe (2004) volume, provided further support for the arguments emerging from earlier work in the field: that washback is not easy to predict or control, and that the shape it assumes is influenced not only by tests but by the interaction of numerous factors, including characteristics of the teachers and students involved, characteristics of the educational context and characteristics of the wider social, political and cultural setting. They also reveal the variety of quantitative and qualitative research methods that are now being used to probe the existence of washback and to explain the appearance it takes in different contexts.

A recent survey of washback studies by Spratt (2005) sets out to inform teachers (rather than testers or researchers) 'of the roles they can play and the decisions they can make concerning washback' (5). Spratt looks at approximately 20 studies and discusses what they reveal about washback on the curriculum (e.g. the content of the curriculum, when it is offered, and whether time is taken away from normal teaching to emphasise the areas believed to be assessed in the test), teaching materials (for instance, the use of course books, past papers and other materials), teaching methods (which methods to use and in which measure, the teaching of test-taking skills), feelings and attitudes (which attitudes to promote in learners) and student learning (the appropriacy of student outcomes). She also provides a clear account of the many types of factors now known to influence the nature of washback.

Critical issues and topics

The language testing community has discussed washback for nearly two decades, but there are a number of issues which have yet to be resolved. The first has to do with the difficulty of separating out the influence of tests from the effects of other variables at work in the educational

context. Some of the early references to washback in the language testing field assumed a direct cause and effect relationship between a test and the effects it would have in the classroom. Heaton, for example, wrote that

> If it is a good examination, it will have a useful effect on teaching; if bad, then it will have a damaging effect on teaching.
>
> *(Valette, 1990: 17)*

Research has now shown that the introduction or revision of a test will not automatically cause changes to occur in teaching and learning. While tests may have an influence on attitudes and behaviour, it is not easy to predict how much influence they will have or what this will look like in concrete terms. Even if a test carries high stakes, it is but one factor amongst many that are interacting in any educational setting. It is important to understand all these factors in order to be able to gauge whether the washback intended by the test designers is likely to appear as envisaged or whether it will be diluted or distorted beyond recognition. A good example of how complex the situation can be is provided by Wall (2000, 2005), who compared classroom practice before and after the introduction of a high-stakes test of English in the state school system in Sri Lanka. The test was meant to reflect and reinforce an innovative approach to teaching (new objectives, new content and new teaching methods) embodied in a series of textbooks introduced several years earlier. Observations of classrooms over time and in-depth interviews with teachers after the launch of the test revealed that while some changes had come about in *what* the teachers were teaching (more emphasis on reading and writing, for example), there was little change in *how* they were teaching (more evidence of a word-for-word approach to dealing with reading, for example, than of efforts to develop enabling skills and strategies). Wall's analysis, which was informed by a diffusion-of-innovation framework devised by Henrichsen (1989), showed that the type of teaching that was taking place after the introduction of the test was due not only to the characteristics of the test itself (there were strengths and weaknesses in its design), but also to conditions in place in the educational setting before the test was introduced (e.g. traditional ways of teaching) and factors at play when the test was introduced and for some time thereafter (e.g. inadequate teacher support, and problems in communication and resourcing).

Later studies have also emphasised the importance of looking at factors beyond the test when attempting to predict what form washback will take in new surroundings. Green (2007: 25), for example, points to the need to take the test users' characteristics into account:

> Differences among participants in the perceptions of test importance and difficulty, and in their ability to accommodate to test demands, will moderate the strength of any effect, and, perhaps, the evaluation of its direction.

Spratt (2005: 29) presents a long list of factors which have been shown to influence a test's washback. These include teacher-related factors (beliefs, attitudes, education and training), resourcing (with a focus on teacher-made and commercial materials), the conditions at the school where teaching and test preparation is taking place and the attributes of the test in question (e.g. its proximity, its purpose, the status of the language it tests, the formats it employs, the weighting of the different sections and how familiar the test is to teachers). Unfortunately, however, the fact that we are now more aware of the complexity of washback does not necessarily lead to greater success in predicting its coverage or intensity.

The second issue has to do with the dilemma many teachers face when they are preparing their students for a test that has important consequences. Most teachers work in a system with a

curriculum that is to be honoured and a syllabus that should be adhered to. If a test is introduced which is known (or believed) to sample only some of the syllabus then the teacher must decide whether to cover the whole syllabus or whether to 'teach to the test'. When educators such as Popham promoted the idea of 'measurement-driven instruction' they envisaged a system where tests were 'properly conceived and implemented' (Valette, 1987: 680). The term 'teaching to the test' implies that the test is not so conceived, and that the teachers are 'doing something in teaching that may not be compatible with (their) own values and goals, or with the values and goals of the instructional program' (Bachman and Palmer, 2010: 108). Teachers may well understand that they are not serving their students' long-term interests by narrowing their teaching to address what they perceive to be the demands of the test. However, they are also likely to feel pressured to ensure that their students perform well, and this can lead to what Spratt (2005: 24) calls 'a tension between pedagogical and ethical decisions'. The most effective solution to this problem is for test designers to 'sample widely and unpredictably' (Hughes, 2003: 54), in order to encourage teachers to teach all the points in the syllabus. Unless this occurs, in teachers will still be faced with hard decisions, and some may find themselves, without devious intentions, engaging in the sort of activities that researchers such as Haladyna *et al.* (1991) might see as less than ethical.

The third issue has to do with the responsibilities of test developers with regard to the washback of their tests and/or any impact that extends beyond the classroom, into the educational system or even greater society. If the test is having a beneficial influence then there is no problem, but if teachers or institutions narrow their teaching too drastically or if there are other unintended consequences should the blame be placed on the test developers? In Messick's view

> washback is a consequence of testing that bears on validity only if it can be evidentially shown to be an effect of the test and not of other forces operative on the educational scene.
>
> *(Valette, 1996: 242)*

As stated in 'Historical perspectives' above, Messick sought to minimise construct under-representation and construct-irrelevant difficulty, so it follows logically that the test designer would be responsible for negative consequences if the test design suffered from either of these sources of invalidity. It is more difficult to assign responsibility in cases where unintended consequences stem not from poor test design but from the misuse of tests, when tests are used for purposes they were not intended for (see Davies, this volume). An example of such unintended (and generally considered to be negative) consequences would be the exclusion of immigrants from a country as a result of low scores on a language test not designed for this type of screening. The Standards for Educational and Psychological Testing state that 'If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use' (AERA/APA/NCME, 1999 Standard 1.4: 28); however, there is much debate in the language testing community about how much warning test developers should give to the users and what other actions they might take to prevent their tests from being used inappropriately.

Current contributions and research

As can be seen in the preceding sections, the study of washback has expanded greatly over a relatively short period of time. Among the most recent areas to be explored are those of teacher beliefs, washback on learners and washback.

Spratt (2005) lists the types of teacher beliefs that have been seen as important mediators of washback. These includes beliefs about

- the reliability and fairness of the exam
- what constitutes effective teaching methods
- how much the exam contravenes their current teaching practices
- the stakes and usefulness of the exam
- their teaching philosophy
- the relationship between the exam and the textbooks, and
- their students' beliefs.

Only recently, however, have attempts been made to systematically study the nature of teachers' beliefs and the influence that these can have on their classroom behaviour. Huang (2009) used insights from social psychology, in particular the Theory of Planned Behaviour (Ajzen, 2006), to analyse data she had obtained through teacher diaries, observations, and in-depth interviews. Her purpose was to distinguish between teachers' behavioural beliefs (what they believed about the four types of teaching behaviour that a new high-stakes test was meant to encourage), their normative beliefs (what they believed important people around them – district inspectors, their head teacher, their peers, their students and the students' parents – expected of them with regard to the behaviours) and their control beliefs (what they believed was achievable given the possibilities and the constraints of the situation they were working in). Huang showed how the interaction between these different types of beliefs led to intentions about how to how to teach, and how these intentions were modified by actual circumstances before being transformed into observable behaviour. Huang's study provides a good example of how insights from other disciplines can contribute to our understanding of important concepts in language assessment. (See also Burrows, 2004, for an analysis of teachers according to whether they are adopters, adapters or resisters to innovation, and how this type of categorisation allows predictions of how they will react to changes in assessment policy.)

Another relatively new focus of study has to do with the washback of tests on learners, as seen through the eyes of learners. Washback studies have often been concerned with how tests influence teachers, perhaps because of the central role that teachers play in the classroom but also because of the practical difficulties of investigating student attitudes and behaviour. Notable exceptions include Gosa (2004) and Tsagari (2009), who looked at students' reactions to tests as recorded in their diaries, and Watanabe (2001), who used interviews to explore the ways in which tests motivated learners. A recent and innovative approach was described in Huhta *et al.* (2006), who studied a group of students for four months, from the time they started preparing for a high-stakes national school test to after they received their results. The researchers asked the students to keep an oral diary as they prepared for the test, and got them to discuss their reactions with fellow students after they had taken it. The data were analysed using a type of discourse analysis from the field of social psychology, focusing on

> how stakeholders (e.g. test-takers or raters) come to talk or write about a test in attempts to make sense of taking it or rating it … at the same time evaluating aspects related to it, including actions, thoughts and feelings.
>
> *(Huhta et al., 2006: 331)*

The researchers were particularly interested in the roles that the students adopted as they talked about their relationship with the test – whether they saw or portrayed themselves as hardworking or lazy, well or poorly skilled, lucky or unlucky, or 'cool' or nervous. They found that the students adopted different roles at different stages of the report period, and sometimes even changed roles during the same situation. They concluded that the relationship between any

student and a test is an individual and sometimes emotional one. This suggests that studies which talk about washback to 'the learners' may be simplifying matters unduly. (See also Ferman, 2004, for a detailed account of learner test preparation strategies.)

A third new area of interest concerns the detailed tracking of washback over a long period of time. Longitudinal studies are not in themselves new. Shohamy *et al.* (1996), for example, compared the washback occurring soon after the introduction of two national language tests in Israel and the washback of the same tests a number of years later. They found that there were differences in the way the washback of the two tests had developed, and that these were due to differences in the purposes of the tests, the status of the languages involved and features having to do with the tests' internal structure. Such 'before and after' studies have been helpful in shaping our understanding of how washback can change from one period to another, but what they have not given us is an understanding of how it is developing at various points along the way. Wall and Horák (2011) report on a five-year study of the impact of changes in the TOEFL on teaching in a sample of test-preparation institutions in different countries in Europe. The study was divided into four phases: a baseline study which looked at teaching practices before the teachers and institutions involved were aware that a new test was on the horizon; a 'transition phase' which tracked the teachers' reactions as they began learning about changes in the test and had to think about redesigning their preparation courses; a further transition phase, which analysed the course books that the teachers had selected and the use they made of them as they planned and delivered their early redesigned courses; and a final phase which looked at the teachers' classes once the new test had had some time to 'settle' in their countries and compared the teaching then with the teaching that had been taking place in the baseline phase. The unique feature of this study was that it traced several teachers from the start to the finish of the project, probing their awareness and understanding of the new (and old) tests, their developing abilities to cope with the new test demands, their questions about how best to help their students, and their decisions about the materials and teaching methods that would be most appropriate for their own circumstances. The continuous and in-depth communication with the teachers allowed the researchers to feel confident in their claim that 'an evidential link' had been established between the test and the teaching practice in place after its introduction. Although there are considerable resourcing demands in this kind of project there is also the potential to feed information back to the testing agency on a regular basis, and thereby to enable adjustments in communication about the test and in arrangements for teacher support.

Main research methods

Hawkey (2006) reminds us that if one of the aims of a test development project is to create positive impact (in our case, positive washback), then a key part of the process of test validation is to investigate whether the test has achieved its intended purpose. Hawkey states that the purpose of such studies is

> to measure and analyse both outcomes, for example test results or subsequent performance on the criteria the test is measuring, and processes, for example the learning and teaching approaches and activities of programmes preparing candidates for a test.
>
> *(Hawkey, 2006: 21)*

It is important to add that a further product of such studies should be an understanding of why the test has *not* produced the desired washback, so that this information can be fed back into the test design process and/or the testing agency's communication with the test users, in particular with teachers and learners.

Bailey (1996) questions the notion of 'measuring' washback, since this term is associated with experimental research and 'identifying, operationally defining and controlling the variables that impinge upon the desired measurements' (272). Bailey stresses the difficulty of isolating washback from other variables that may affect teaching and learning, and the fact that washback must be studied in naturally occurring settings (rather than laboratory conditions), with a non-random sample of subjects. Although some washback outcomes (e.g. test results) can be measured, other outcomes and most processes can only be analysed and described. Bailey advises that such analysis should involve methodological 'triangulation' – looking at the situation from various angles, using, for example, more than one data set, or more than one type of informant or more than one type of method.

Wall and Horák (2007) state that there are two types of washback studies: those that investigate the effects of tests which are already operational (e.g. Green, 2007, which looked at the effects of the IELTS test on teaching in different contexts), and those that investigate the effects of a new test or a test which has been revised substantially. They focus on the second type of study, and describe the procedures that they followed when setting up an investigation to identify changes in classroom practice that may have been brought about by the introduction of a much-revised TOEFL (Wall and Horák, 2011). The essential steps are as follows:

- identify the sorts of washback the test designers originally intended, and the means they planned to use to ensure that this washback actually appeared;
- analyse the new test in detail, to see how it differs from the test it is replacing;
- predict what washback may occur (intended and unintended), using information from the first two steps;
- design a baseline study, to find out what teaching and/or learning look like before participants are aware of the new test demands;
- carry out the baseline study, to document the educational setting and practices before the test is introduced;
- disseminate the results of the study, to inform the test developers of any features in the setting or established practices that might negatively affect the participants' understanding and ability to accommodate the new test demands.

The key component in this kind of investigation is the baseline study, which can serve as a point of comparison when attempts are made after the launch of the test to describe whether it has produced any washback. After carrying out the baseline study researchers need to keep track of the process leading up to the test launch and what happens as the test becomes established in its context. The amount of detail that is possible will be determined by the resources the researchers have access to. The final stage is to describe the teaching and/or learning that are taking place at the end of a time period agreed at the beginning of the project, and then to explain whether any of the features present in the 'after' period can be evidentially linked to the introduction of the test. As stated earlier, it is also important to try to investigate what other factors may be working in the environment to facilitate or hinder the appearance of the desired washback (Wall, 2005).

The process described above is essentially a 'compare and contrast' exercise, usually using the same research instruments and procedures and the same analytical frameworks in both stages of the study. If the purpose of a washback study is to describe the influence of a test which is already embedded in an educational context then it is also necessary to compare and contrast classroom practices. In this case, however, the comparison may be between the 'ordinary' teaching and the test-preparation teaching presented by the same set of teachers, or the teaching that takes place for two or more different tests. Watanabe (2004: 28) presents a detailed scheme of how to organise

such comparisons in order to investigate various kinds of washback. Watanabe also offers advice on general research matters such as how to select a sample of teachers and institutions to participate in a study; how to gain access to them; whether to reveal the purpose of the research; how to organise observations and interviews; how to analyse data; and how to interpret and discuss the results of the analysis.

In addition to deciding on the general approach to a washback study researchers need to consider carefully the specific methods they will use to gather and analyse data. Among methods commonly used are questionnaires and interviews (for teachers, students, principals, inspectors, policy-makers, etc.), focus groups, document analysis (policy documents, test papers, lesson plans, teaching materials and so on), observations, diaries (written or oral, by teachers or learners), and analysis of test scores. Examples of instruments (questionnaires, textbook analysis frameworks, interview protocols, observation forms, and marking rubrics) can be found at the end of most of the case studies in Cheng and Watanabe (2004), and in the appendices of Cheng and Watanabe (2004), Wall (2005), Hawkey (2006), Green (2007), Tsagari (2009) and Wall and Horák (2011).

Future directions

This survey of the literature on washback shows that a great deal of energy and effort has gone into the study of this phenomenon since the first critical questioning began in the early 1990s. Considerable discussion has taken place about the nature of washback, the factors that can facilitate or hinder its appearance and intensity, and the steps that testing agencies should take to ensure that the effects of their tests are viewed as positive by teachers, learners and others who are affected by test results and the use made of them. Numerous case studies have been undertaken which have commented in detail on the characteristics of particular tests and how they have or have not produced the washback their designers originally intended. While it is encouraging to see the amount of interest washback has generated over the years, it can also be frustrating to see how many directions the research has gone off into and how little connection there seems to be between researchers working on the same types of questions. Although most researchers include references to 'classic' washback studies and will review more recent studies if they happen to relate to the theme they are interested in, it is rare to see research that truly builds on work that has been done previously, which replicates or only slightly adapts the work of others. The washback literature is full of interesting individual studies but it seems to lack an overall coherence. This may be inevitable, given the need for researchers to display originality in order to gain their doctorates or to have their articles accepted by prestigious journals. It may also be a function of the relative newness of this area in the field of language assessment – while 20 years can seem a long time in some ways, it is still only 20 years.

Cheng (2008) expresses similar reservations about the coherence of the washback literature, and she argues that while individual studies of different tests in different contexts have their value,

> [i]t would be the best use of resources if a group of researchers could work collaboratively and cooperatively to carry out a series of studies around the same test within the same educational context.
>
> *(Cheng, 2008: 360)*

Cheng envisages a situation where researchers would look at different aspects of the same situation and would be able to cross-reference their findings to build up a comprehensive picture of the washback operating therein. Spratt (2005) seems to be thinking along the same lines, although she

advocates using 'parallel methodologies' in different contexts, to investigate 'some of the apparent contradictions in the findings to date'.

I concur. I would hope to see a consolidation of the many research findings we already have in the near future, so that we can produce confident statements about how tests affect participants, processes and products (especially learning outcomes) over the next 20 years. My personal interests are in helping testing agencies to create the best possible washback with the least waste of anyone's effort, and in helping researchers to understand how to design the most effective investigations for the tests and the contexts they wish to study. I would therefore also hope to see two very practical outcomes from the collaboration of many researchers: a set of evidence-based guidelines for creating positive washback, and a set of authoritative but user-friendly guidelines on how to design a washback study.

Further reading

Chapman, D. W. and Snyder, C. W. (2000). Can high-stakes national testing improve instruction: Re-examining conventional wisdom. *International Journal of Educational Development* 20: 457–74. Chapman and Snyder review five propositions for how to use high-stakes national testing to improve classroom teaching practices and student learning. They argue that such propositions often fail because policy-makers do not understand the links that must be present (e.g. increased resources to low-achieving schools, teacher and parent concern about poor scores, parent and community pressure on teachers to improve their practices) to convert strategy into the desired outcomes. A model of these linkages is presented to guide future policy-makers in their attempts to improve teaching through testing.

Rea-Dickins, P. and Scott, C. (2007). Washback from language tests on teaching, learning and policy: Evidence from diverse settings. *Assessment in Education: Principles, Policy & Practice* 14: 1–7. Rea-Dickins and Scott provide an overview of the notion of washback and associated issues in this editorial for a special issue of *Assessment in Education*. They summarise six papers commissioned for the issue, which include investigations into washback in settings which differ not only in terms of the groups being studied but also geographically. The authors/editors conclude that washback should be re-defined as a 'context-specific shifting process, unstable, involving changing behaviours in ways that are difficult to predict'.

References

AERA/APA/NCME (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Airasian, P. (1988). Measurement-Driven Instruction: A Closer Look. *Educational Measurement: Issues and Practice* Winter: 6–11.

Ajzen, I. (2006). Constructing a TpB Questionnaire: Conceptual and Methodological Considerations. www.people.umass.edu/aizen/pdf/tpb.measurement.pdf (accessed 4/1/11).

Alderson, J. C. and Hamp-Lyons, L. (1996). TOEFL preparation courses: a study of washback. *Language Testing* 13: 280–97.

Alderson, J. C. and Wall, D. (1993). Does Washback Exist? *Applied Linguistics* 14: 115–29.

Andrews, S. (2004). Washback and Curriculum Innovation. In L. Cheng and Y. Watanabe (eds), *Washback in Language Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 37–50.

Bachman, L. and Palmer, A. (2010). *Language Assessment in Practice*. Oxford, UK: Oxford University Press.

Bailey, K. (1996). Working for washback: a review of the washback concept in language testing. *Language Testing* 13: 257–79.

—— (1999) *Washback in Language Testing*. TOEFL Monograph No. MS-15. Princeton, NJ: Educational Testing Service.

Burrows, C. (2004). Washback in classroom-based assessment: a study of the washback effect in the Australian adult migrant English Program. In L. Cheng and Y. Watanabe (eds), *Washback in Language Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 113–28.

Cheng, L. (2005). *Changing Language Teaching Through Language Testing: a washback study*. Cambridge, UK: Cambridge University Press and Cambridge ESOL.

—— (2008). Washback, Impact and Consequences. In E. Shohamy and N. H. Hornberger (eds), *Encyclopedia of Language and Education*, 2nd edn. Vol. 7. *Language Testing and Assessment*. New York, NY: Springer Science+Business Media LLC, 349–64.

Cheng, L. and Curtis, A. (2004). Washback or backwash: a review of the impact of testing on teaching and learning. In L. Cheng and Y. Watanabe (eds), *Washback in Language Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 3–17.

Cheng, L. and Watanabe, Y. (eds), (2004). *Washback in Language Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Eckstein, M.A. and Noah, H. J. (1993). *Secondary School Examinations: international perspectives on policies and practice*. New Haven, CT: Yale University Press.

Ferman, I. (2004). The washback of an EFL national oral matriculation test to teaching and learning. In L. Cheng and Y. Watanabe (eds), *Washback in Language Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 191–210.

Fredericksen, J.R. and Collins, A. (1989). A systems approach to educational testing. *Educational Researcher* 18: 27–32.

Fullan, M. G. and Stiegelbauer, S. (1991). *The New Meaning of Educational Change*, 2nd edn. London, UK: Cassell Educational Limited.

Gosa, C. M. C. (2004). *Investigating washback: a case study using student diaries*. Unpublished PhD Thesis, Lancaster University, UK.

Green, A. (2007). *IELTS Washback in Context: preparation for academic writing in higher education*. Cambridge, UK: Cambridge University Press and Cambridge ESOL.

Haladyna, T. M., Nolen, S. B. and Haas, N.S. (1991). Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution. *Educational Researcher* 20: 2–7.

Hawkey, R. (2006). *Impact Theory and Practice: Studies of the IELTS Test and Progetto Lingue 2000*. Cambridge, UK: Cambridge University Press and Cambridge ESOL.

Heaton, J. B. (1990). *Classroom Testing*. Harlow, UK: Longman.

Henrichsen, L. E. (1989). *Diffusion of Innovations in English Language Teaching: The ELEC Effort in Japan, 1956–1968*. New York, NY: Greenwood Press.

Heyneman, S. P. and Ransom, A.W. (1990). Using examinations and testing to improve educational quality. *Educational Policy* 4: 177–92.

Huang, L. (2009). *Washback on Teachers' Beliefs and Behaviour: Investigating the Process from a Social Psychology Perspective*. Unpublished PhD Thesis, Lancaster University.

Hughes, A. (1994). *Backwash and TOEFL 2000*. Unpublished manuscript, commissioned by Educational Testing Service.

—— (2003). *Testing for Language Teachers*, 2nd edn. Cambridge, UK: Cambridge University Press.

Huhta, A., Kalaja, P. and Pitkänen-Huhta, A. (2006). Discursive construction of a high-stakes test: the many faces of a test-taker. *Language Testing* 23: 326–50.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (ed.), *Critical Issues in Curriculum: eighty-seventh yearbook of the National Society for the Study of Education*. Chicago, IL: University of Chicago Press, 83–121.

Mehrens, W. A. and Kaminsky, J. (1989). Methods for Improving Standardized Test Scores: Fruitful, Fruitless or Fraudulent? *Educational Measurement: Issues and Practices* 8: 14–22.

Messick, S. (1996). Validity and washback in language testing. *Language Testing* 13: 241–56.

Pearson, I. (1988). Tests as levers for change. In D. Chamberlain and R. Baumgardner (eds), *ESP in the Classroom: practice and evaluation*. ELT Documents 128. London, UK: Modern English Publications, 98–107.

Popham, W. J. (1987). The merits of measurement-driven Instruction. *Phi Delta Kappan* 68: 679–82.

Shohamy, E., Donitsa-Schmidt S. and Ferman, I. (1996). Test impact revisited: washback effect over time. *Language Testing* 13: 298–317.

Spratt, M. (2005). Washback and the classroom: the implications for teaching and learning of studies of washback from exams. *Language Teaching Research* 9: 5–29.

Tsagari, D. (2009). *The Complexity of Test Washback*. Frankfurt, Germany: Peter Lang.

Wall, D. (1996). Introducing New Tests into Traditional Systems: Insights from General Education and from Innovation Theory. *Language Testing* 13, 3: 334–54.

—— (2000). The impact of high-stakes testing on teaching and learning: can this be predicted or controlled? *System* 28: 499–509.

—— (2005). *The Impact of High-Stakes Examinations on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory*. Cambridge, UK: Cambridge University Press and Cambridge ESOL.

Wall, D. and Horák, T. (2007). Using baseline studies in the investigation of test impact. *Assessment in Education* 14: 99–116.

—— (2011). *The Impact of Changes in the TOEFL on Teaching in a Sample of Countries in Europe*: phase 3, the role of the coursebook, and phase 4, describing change. TOEFL iBT Reports series. Princeton, NJ: Educational Testing Service.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing* 13: 318–33.

—— (2001). Does the university entrance examination motivate learners? A case study of learner interviews. In Akita Association of English Studies (ed.), *Trans-Equator Exchanges: a collection of academic papers in honour of Professor David Ingram*. Akita, Japan: Akita Association of English Studies, 100–10.

—— (2004). Methodology in washback studies. In L. Cheng and Y. Watanabe (eds), *Washback in Language Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 19–36.